(19) JAPANESE PATENT OFFICE

PATENT ABSTRACTS OF JAPAN

(11) Publication number: **06274548 A**

(43) Date of publication of application: **30 . 09 . 94**

(51) Int. Cl **G06F 15/40**
G06F 15/38

(21) Application number: **05061641**

(22) Date of filing: **22 . 03 . 93**

(71) Applicant: **A T R JIDO HONYAKU DENWA KENKYUSHO:KK**

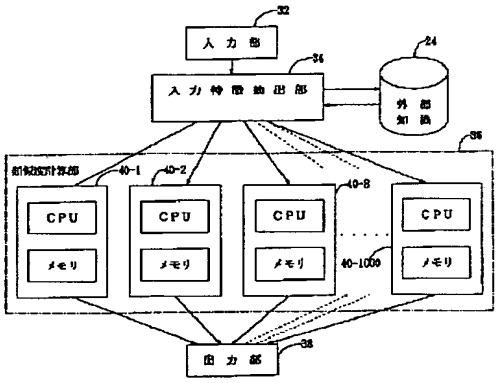
(72) Inventor: **OI KOZO**
SUMIDA EIICHIRO
IIDA HITOSHI

(54) **SIMILARITY DEGREE CALCULATING DEVICE**

(57) Abstract:

PURPOSE: To quickly calculate the degree of similarity between input features acquired from external knowledge and comparison features corresponding to input information and comparison information.

CONSTITUTION: This device is provided with an input part 32 for inputting the input information, the external knowledge 24 for leading features which correspond to supplied information but does not appear in the information itself, similarity degree leading parts 34 and 36 for leading out the features of the input information from the external knowledge 24 and parallelly calculating the degrees of similarity with the respective features of the plural comparison information and an output part 38 for outputting the obtained degrees of similarity.



(19)日本国特許庁（ＪＰ）　(12)**公開特許公報（Ａ）**　(11)特許出願公開番号

特開平6－274548

(43)公開日　平成６年(1994)９月30日

(51)Int.Cl.[5]	識別記号	庁内整理番号	ＦＩ	技術表示箇所
Ｇ０６Ｆ　15/40	５１０　Ｍ	9194－5Ｌ		
15/38	Ｃ	7323－5Ｌ		

審査請求　有　　請求項の数９　ＯＬ　（全 12 頁）

(21)出願番号　特願平5－61641

(22)出願日　平成５年(1993)３月22日

(71)出願人　000127684
株式会社エイ・ティ・アール自動翻訳電話研究所
京都府相楽郡精華町大字乾谷小字三平谷５番地

(72)発明者　大井　耕三
京都府相楽郡精華町大字乾谷小字三平谷５番地　株式会社エイ・ティ・アール自動翻訳電話研究所内

(74)代理人　弁理士　深見　久郎　（外２名）

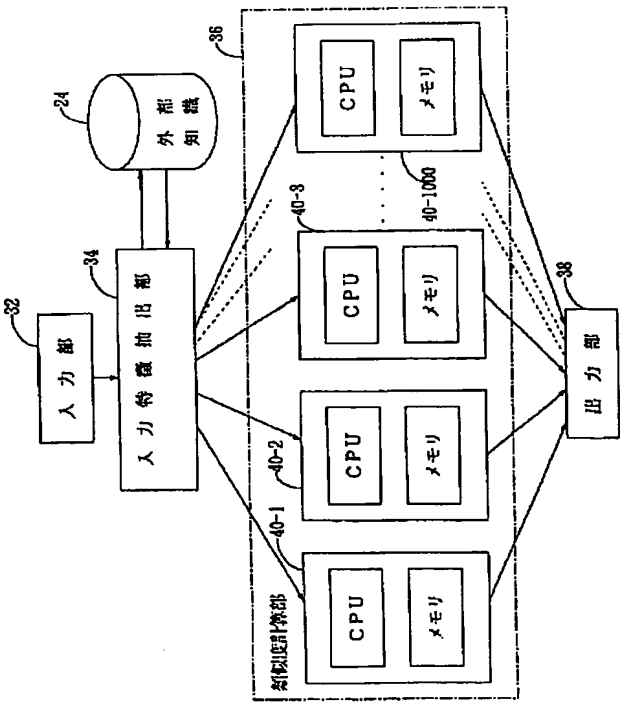
最終頁に続く

(54)【発明の名称】　類似度計算装置

(57)【要約】

【目的】　入力情報と比較情報とに対応して外部知識から獲得した入力特徴および比較特徴の類似度計算を高速化する。

【構成】　入力情報を入力するための入力部３２と、与えられる情報に対応するが情報自体には現われていない特徴を導出するための外部知識２４と、入力情報の特徴を外部知識２４から導出し、複数の比較情報の特徴とのそれぞれと、類似度計算を並列に行なうための類似度導出部３４、３６と、得られた類似度を出力するための出力部３８とを含む。



【特許請求の範囲】

【請求項１】　複数個の予め定める比較情報との間での類似度を計算する対象となる情報を入力するための手段と、

与えられる情報に対応するが、前記与えられる情報自体には現われていない特徴を導出するための外部知識を格納するための手段と、

前記複数個の比較情報に対応して前記外部知識を用いて導出される前記比較情報の特徴のそれぞれと、前記入力手段を介して入力される情報に対応して前記格納手段から導出される特徴との間で所定の類似度計算を並列に行なって、前記比較情報の特徴のそれぞれと、前記入力される情報の特徴との類似度を導出するための手段と、

前記導出された類似度を出力するための手段とを含む、類似度計算装置。

【請求項２】　前記類似度を導出するための手段は、前記複数個の比較情報に対応して前記外部知識を用いて動的に導出される前記比較情報の特徴のそれぞれと、前記入力手段を介して入力される情報に対応して前記格納手段から動的に導出される特徴との間で所定の類似度計算を並列に行なって、各前記比較情報の特徴と前記入力される情報の特徴との間での類似度を導出するための手段を含む、請求項１に記載の類似度計算装置。

【請求項３】　前記外部知識から導出される特徴がｎ桁のｍ進数で表現されることを特徴とし、ｎおよびｍはともに自然数である、請求項１に記載の類似度計算装置。

【請求項４】　前記類似度を導出するための手段が、前記入力される情報について導出される特徴の各桁の文字と、各前記比較情報について導出される特徴の対応する各桁の文字とを比較し、その比較の結果に従って類似度を計算するための手段を含む、請求項３に記載の類似度計算装置。

【請求項５】　前記類似度を計算するための手段が、前記比較の結果相互に一致する文字の数に従って類似度を計算するための手段を含む、請求項４に記載の類似度計算装置。

【請求項６】　前記類似度を導出するための手段が、前記複数個の比較情報のうちの予め定める１つに対応して前記外部知識を用いて導出される特徴と、前記入力手段を介して与えられる情報に対応して前記格納手段から導出される特徴との間で所定の類似度計算を行なって、前記複数個の比較情報の前記予め定める１つの特徴と前記入力される情報の特徴との間での類似度を導出するための、複数個の、かつ相互に並列に動作可能な演算手段を含む、請求項１に記載の類似度計算装置。

【請求項７】　各前記演算手段が電子計算機を含む、請求項６に記載の類似度計算装置。

【請求項８】　前記複数個の電子計算機が、ネットワークを介して前記入力特徴抽出部と前記出力部とに接続される、請求項７に記載の類似度計算装置。

【請求項９】　前記類似度を導出するための手段が、前記比較情報を格納する、格納内容で参照可能な記憶手段と、

前記入力される情報の特徴の、所定の１または複数個の桁の文字と一致する文字を、前記１または複数個の桁と一致する位置に有する前記比較情報の特徴を、前記格納内容で参照可能な記憶手段において検出するための手段と、

前記検出手段による検出結果に基づいて各前記比較情報の特徴と前記入力される情報の特徴との間の類似度を計算するための手段を含む、請求項３に記載の類似度計算装置。

【発明の詳細な説明】

【０００１】

【産業上の利用分野】この発明は情報処理分野、特に機械翻訳や情報検索や質問応答などの自然言語処理、画像処理、音声処理などの分野に関し、特に、入力されるある情報（以下「入力情報」と呼ぶ）と、この入力情報との比較が行なわれる他の情報（以下「比較情報」と呼ぶ）のそれぞれとの間の類似度を計算する装置に関する。

【０００２】

【従来の技術】従来、入力情報と比較情報のそれぞれとの間の類似度の計算を行なう類似度計算装置では、計算は入力情報および比較情報自体をデータとして行なわれていた。

【０００３】たとえば、画像検索について述べる。画像検索とは、比較情報として予め準備されている大量の画像の中から、入力画像に類似した画像を検索する処理である。この場合、１つの入力画像と１つの比較画像との間の類似度を求めるときには、各画像の各ドットの色および濃淡のデータ、すなわち、画像そのもののデータを用いていた。

【０００４】これと対照的に、入力情報には現われていない特徴を、外部知識から獲得して類似度計算を行なう事例もある。以下、特に機械翻訳の事例について説明する。

【０００５】機械翻訳システムの１タイプとして、大量の用例（原文と訳文との対）と入力文とを比較し、入力文に最も類似した用例を求めてそれを基に訳文を生成するタイプのものがある。そのようなタイプの機械翻訳システムにおいては、入力文すなわち入力情報と用例すなわち比較情報との間の類似度は、その入力文における単語（以下「入力単語」と称する）と用例における単語（以下「比較単語」と称する）との間の類似度を、たとえば類語辞書のような外部知識から獲得した情報を用いて計算する。この類語辞書のような外部知識から獲得した情報は、入力単語あるいは比較単語にそれぞれ対応したものではあるが、入力単語あるいは比較単語には現われていないような特徴に関する情報である。

【０００６】このように、外部知識から獲得した特徴に関する情報を用いて類似度計算を行なう装置は、従来、入力情報と比較情報のそれぞれとの間の類似度計算を逐次処理で行なっていた。

【０００７】図１２は、従来のこのような類似度計算装置において、入力情報と比較情報のそれぞれとの間の類似度計算をする際の手順を示すフローチャートである。

【０００８】たとえば、電子計算機を使用して類似度計算を行なう場合、比較情報に対応し比較情報には現われていない比較特徴は予め計算機のメモリ上に格納されている。そして類似度計算は次のような手順によって行なわれている。

【０００９】図１２を参照して、まずステップＳ００１で、入力情報に対応し入力情報には現われていない入力特徴を外部知識から抽出する処理が行なわれる。

【００１０】続いてステップＳ００２～Ｓ００５で、予め計算機のメモリ上に格納されていた比較特徴を１つずつＣＰＵ（中央演算処理装置）内に取り出して、順次入力情報の特徴との間の類似度計算を行なっていた。

【００１１】

【発明が解決しようとする課題】しかし、入力情報と比較情報とに対応し、入力情報あるいは比較情報には現われていない特徴を用いて類似度計算をする場合には、処理時間が非常にかかるという問題点がある。たとえば、入力文と大量の用例とを比較し、最も類似した用例を求めてそれを基に訳文を生成するタイプの機械翻訳システムにおいて、より正確な翻訳を行なおうとすれば用例数をより多くすることが必要である。ところが、数万の用例を用いて類似度計算を行なう場合には、短い文でも入力文とすべての用例との間の類似度の計算に数分を要する。そのためにこのようなタイプの機械翻訳システムでリアルタイム処理を実現することは不可能であった。

【００１２】それゆえに請求項１～９に記載の発明の目的は、入力情報と比較情報のそれぞれとの間の類似度を、外部知識から獲得した、入力情報に対応した入力特徴および比較情報に対応した比較特徴を用いて求める類似度計算において、計算速度を大幅に向上できる類似度計算装置を提供することである。

【００１３】

【課題を解決するための手段】請求項１に記載の類似度計算装置は、図１に示されるように、複数個の予め定める比較情報との間での類似度を計算する対象となる情報を入力するための手段２０と、与えられる情報に対応するが、前記与えられる情報自体には現われていない特徴を導出するための外部知識を格納するための手段２４と、前記複数個の比較情報に対応して前記外部知識を用いて導出される前記比較情報の特徴のそれぞれと、前記入力手段を介して入力される情報に対応して前記格納手段から導出される特徴との間で所定の類似度計算を並列に行なって、前記比較情報の特徴のそれぞれと、前記入力される情報の特徴との類似度を導出するための手段２２と、前記導出された類似度を出力するための手段２６とを含むことを特徴とする。

【００１４】請求項２に記載の類似度計算装置は請求項１に記載の装置であって、類似度を導出するための手段は、複数個の比較情報に対応して外部知識を用いて動的に導出される比較情報の特徴のそれぞれと、入力手段を介して入力される情報に対応して格納手段から動的に導出される特徴との間で所定の類似度計算を並列に行なって、各比較情報の特徴と入力される情報の特徴との間での類似度を導出するための手段を含む。

【００１５】請求項３に記載の類似度計算装置は請求項１に記載の装置であって、外部知識から導出される特徴がｎ桁のｍ進数（ｎおよびｍはともに自然数）で表現されることを特徴とする。

【００１６】請求項４に記載の類似度計算装置は請求項３に記載の装置であって、類似度を導出するための手段が、入力される情報について導出される特徴の各桁の文字と、各比較情報について導出される特徴の対応する各桁の文字とを比較し、その比較の結果に従って類似度を計算するための手段を含むことを特徴とする。

【００１７】請求項５に記載の類似度計算装置は請求項４に記載の装置であって、類似度を計算するための手段が、比較の結果相互に一致する文字の数に従って類似度を計算するための手段を含むことを特徴とする。

【００１８】請求項６に記載の類似度計算装置は請求項１に記載の装置であって、類似度を導出するための手段が、複数個の比較情報のうちの予め定める１つに対応して外部知識を用いて導出される特徴と、入力手段を介して与えられる情報に対応して格納手段から導出される特徴との間で所定の類似度計算を行なって、複数個の比較情報の予め定める１つの特徴と、入力される情報の特徴との間での類似度を導出するための、複数個の、かつ相互に並列に動作可能な演算手段を含む。

【００１９】請求項７に記載の類似度計算装置は請求項６に記載の装置であって、各演算手段は電子計算機を含む。

【００２０】請求項８に記載の類似度計算装置は請求項７に記載の装置であって、複数個の電子計算機が、ネットワークを介して入力手段と外部知識を格納するための手段と出力部とに接続される。

【００２１】請求項９に記載の類似度計算装置は請求項３に記載の装置であって、類似度を導出するための手段が、比較情報を格納する、格納内容で参照可能な記憶手段と、入力される情報の特徴の、所定の１または複数個の桁の文字と一致する文字を、１または複数個の桁と一致する位置に有する比較情報の特徴を、格納内容で参照可能な記憶手段内において検出するための手段と、検出手段による検出結果に基づいて各比較情報の特徴と入力される情報の特徴との間の類似度を計算するための手段

とを含む。

【００２２】

【作用】請求項１～９に記載の類似度計算装置においては、入力情報と比較情報のそれぞれとの間の類似度が、外部知識から獲得した、入力情報に対応した入力特徴および比較情報に対応した比較特徴を用いて、並列に計算される。したがって従来の逐次計算処理よりも高速に類似度計算を行なうことができる。

【００２３】並列演算を実現するためには、請求項６、７に記載のように、複数個の、相互に並列に動作可能な電子計算機などの演算手段を用いることができる。これら電子計算機は、請求項８に記載のようにネットワークを介して接続されてもよい。また請求項９に記載のように、格納内容で参照可能な記憶手段を用いて類似度の計算を並列的に実行してもよい。

【００２４】

【実施例】以下、本発明の３つの実施例を順に説明する。３つの実施例とも、１つの入力情報と１０００個の比較情報との間の類似度計算を行なう装置である。情報としては単語を用い、入力された単語（以下「入力単語」と称する）と、比較対象となる単語（以下「比較単語」と称する）１０００語との間の類似度計算を行なう例を説明する。

【００２５】なお、以下の説明では単語の間の類似度計算を行なう場合を例として本発明を説明するが、本発明は単語の比較のみに限定されるわけではなく、前述のような画像情報や、音声情報などにおける類似度計算に対しても適用可能である。

【００２６】実施例の説明をする前に、これら実施例が使用される背景について説明する。前述の、入力文と大量の用例（原文と訳文との対）とを比較して最も類似した用例を求めてそれを基に訳文を生成するタイプの機械翻訳システムについて考える。このタイプの機械翻訳システムにおいては、入力文および用例を構成する各単語ごとの類似度を計算することによって最も類似した用例が求められる。そしてその用例を基に訳文が生成される。

【００２７】単語の類似度は以下のようにして求められる。まず、外部知識としての類語辞書を準備する。この類語辞書から、入力単語または比較単語に対応した特徴に関する情報として「類語コード」を獲得し、その類似コードを用いて単語間の類似度を求めている。

【００２８】類語辞書とは、単語の意味を階層的に分類した体系に基づき、各単語に類語コードを付与したものを格納した辞書である。図２に、４階層からなる類語辞書の分類体系の一部を示す。

【００２９】図２を参照して、各単語に付与される類語コードは３桁の１０進数からなる。類語コードの１００の位、１０の位、１の位はそれぞれ、分類体系の大分類、中分類、小分類を表わす。図２に示される例において、単語「取材」は、大分類が［取引］、中分類が［報道］、小分類が［編集］である分類に属し、その分類の類語コードは４５７となっている。

【００３０】なお、本実施例の説明においては、図２に示すように３桁の１０進数の類語コードを用いている。しかし、本発明はこのような類語コードではなく、一般的なｎ桁ｍ進数やベクトルなど、様々な形の情報を類語コードと同様のものとして処理可能である。

【００３１】（１）　第１の実施例

図３を参照して、本発明の第１の実施例の類似度計算装置は、入力部３２と、入力特徴抽出部３４と、外部知識２４と、類似度計算部３６と、出力部３８とを含む。

【００３２】入力部３２はキーボード、文字認識装置、音声認識装置などからなる。入力部３２は、入力単語を入力特徴抽出部３４に与えるためのものである。

【００３３】外部知識２４は、ハードディスクやメモリなどからなる。この外部知識２４は、図４に示されるような、各単語の見出しとその類語コードが対になったデータを多数格納している。

【００３４】入力特徴抽出部３４は、入力部３２から与えられる入力単語に対応した類語コード（以下「入力コード」と称する）を外部知識２４から抽出し、類似度計算部３６に与えるためのものである。

【００３５】類似度計算部３６は、１０００個のコンピュータ４０－１～４０－１０００を含む。各コンピュータ４０－１～４０－１０００の入力は入力特徴抽出部３４の出力に接続されている。また各コンピュータ４０－１～４０－１０００の出力は出力部３８の入力に接続されている。これらコンピュータ４０－１～４０－１０００の各々は、ＣＰＵとメモリとを含む。１つのコンピュータのメモリには、１つの比較単語に対応した類語コード（以下「比較コード」と称する）および類似度計算のためのプログラムが格納されている。これらコンピュータ４０－１～４０－１０００には入力特徴抽出部３４から入力コードが与えられる。各コンピュータ４０－１～４０－１０００は、与えられた入力コードと各コンピュータのメモリに格納されている比較コードとの間の類似度を後述するような方法に従って計算し、計算された類似度を出力部３８に与える。これらコンピュータ４０－１～４０－１０００の各メモリに格納されている比較コードの一例が図５に示されている。図５においてたとえばコンピュータ４０－１に格納されている比較コード「４２６」は、図４に示されるように比較単語「販売」に対応する類語コードである。他の比較コードも同様に図４に示される外部知識２４内の或る単語に対応する類語コードとなっている。

【００３６】出力部３８は、表示装置、印刷装置などからなる。出力部３８は、類似度計算部３６のコンピュータ４０－１～４０－１０００から与えられた類似度を出力するためのものである。

【００３７】図６は、図３のコンピュータ４０－１～４０－１０００の各々で行なわれる、入力コードと比較コードとの間の類似度の求め方を示す。図６において最左欄入力コードと比較コードとの間に成立する条件を示す。中欄は、最左欄に示される条件に適合したときの類似度を示す。最右欄は、最左欄に示される条件に適合するような入力コードと比較コードとの対を示す。

【００３８】入力コードと比較コードとの条件欄における記号「Ｉ１」「Ｉ２」「Ｉ３」はそれぞれ、入力コードの１００の位と、１０の位と、１の位とを表わす。記号「Ｃ１」「Ｃ２」「Ｃ３」はそれぞれ、比較コードの１００の位と、１０の位と、１の位とを表わす。

【００３９】図６を参照して、第２行目で示されるように、入力コードと比較コードとが、１００の位と、１０の位と、１の位とのいずれでも一致する場合には、類似度は３となる。第３行目に示されるように１００の位と１０の位とが一致し、１の位のみが異なる場合には類似度は２となる。１００の位のみが一致し、１０の位が異なる場合には、図６の４行目に示されるように類似度は１となる。図６の第５行目に示されるように入力コードと比較コードの１００の位が互いに異なる場合には類似度は０となる。

【００４０】以下、図３に示される類似度計算装置の動作を、入力単語が「取材」である場合を例にして説明する。入力単語「取材」が入力部３２を介して入力されると、その入力単語「取材」は入力特徴抽出部３４に与えられる。

【００４１】入力特徴抽出部３４は、入力単語「取材」に対応した類語コード（入力コード）を外部知識２４から抽出する。この場合図４の第４行目に示されるように、入力単語「取材」に対応した類語コードは４５７となっているので、入力コードとして４５７が得られる。この入力コード４５７は、類似度計算部３６のコンピュータ４０－１～４０－１０００のすべてに与えられる。

【００４２】各コンピュータ４０－１～４０－１０００は、入力コード「４５７」が与えられると、それぞれ独立に類似度計算を行なう。各コンピュータは、入力コード「４５７」と、そのコンピュータに割り当てられている比較コードとを比較し、図６に示される類似度の求め方に従って類似度を求める。

【００４３】たとえば図５を参照して、コンピュータ４０－１のメモリに格納されている比較コードは「４２６」である。したがって図６の第４行目の条件（ハ）により類似度１となる。コンピュータ４０－２では、比較コードは「１４９」である。条件（ニ）により類似度０となる。コンピュータ４０－３では、比較コードは「４５８」である。条件（ロ）により類似度は２となる。コンピュータ４０－１０００では、比較コードは「７３２」である。条件（ニ）により類似度０となる。他のコンピュータ４０－４～４０－９９９でも同様の類似度計算が行なわれる。類似度計算部３６の各コンピュータ４０－１～４０－１０００は、求められた類似度を出力部３８に与える。

【００４４】出力部３８は、コンピュータ４０－１～４０－１０００からの類似度がすべて与えられると、表示装置や印刷装置などにその類似度を出力する。

【００４５】この第１の実施例では、１０００個の類似度計算がコンピュータ４０－１～４０－１０００により並列に行なわれる。したがって従来の逐次処理による計算に比べて単純計算で約１０００倍高速に行なわれる。

【００４６】（２）　第２の実施例

図７は、本発明の第２の実施例の類似度計算装置の概略構成図である。図７を参照してこの類似度計算装置は、入出力管理部５０と、入出力管理部５０が接続されるネットワーク５２と、ネットワーク５２に接続される１０００個の類似度計算部５４－１～５４－１０００とを含む。

【００４７】入出力管理部５０は、入力部３２と、外部知識２４と、入力特徴抽出部３４と、出力部３８とを含む。図７と図３とにおいて、同一のブロックには同一の参照符号および名称が与えられており、それらの機能も同一である。したがってここではそれらについての詳しい説明は繰返さない。この入出力管理部５０は、ワークステーションやパーソナルコンピュータなどによって構成される。

【００４８】外部知識２４には、図４に示されるデータが格納されている。類似度計算部５４－１～５４－１０００の各々は、ワークステーションやパーソナルコンピュータなどを含む。これら各類似度計算部５４－１～５４－１０００は、ＣＰＵとメモリとを含む。１つの類似度計算部のメモリには、１つの比較単語に対応した類語コード（比較コード）および類似度計算のプログラムが格納されている。

【００４９】以下、この発明の第２の実施例の類似度計算装置の動作を説明する。キーボードなどからなる入力部３２によって入力単語が入力特徴抽出部３４に与えられる。入力特徴抽出部３４は、この入力単語に対応する入力コードを外部知識２４から抽出し、ネットワーク５２を介して類似度計算部５４－１～５４－１０００のすべてに送る。各類似度計算部５４－１～５４－１０００は、送られてきた入力コードと、各類似度計算部のメモリに格納されている比較コードとの間の類似度を第１の実施例における方法と同様の手順で計算する。類似度計算部５４－１～５４－１０００はすべて、求めた類似度を入出力管理部５０にネットワーク５２を介して送る。入出力管理部５０の出力部３８は、前述と同様に表示装置、印刷装置などからなり、類似度計算部５４－１～５４－１０００から送られてきた類似度を出力する。

【００５０】この第２の実施例においても、類似度計算部５４－１～５４－１０００はそれぞれ独立に類似度計

算を行なう。すなわち、１０００個の類似度計算が並列に行なわれる。したがって従来の逐次処理による計算に比べて単純計算で約１０００倍高速に類似度計算が行なわれる。

【００５１】（３）　第３の実施例

図８は、本発明の第３の実施例の類似度計算装置の概略構成を示すブロック図である。図８を参照して、この類似度計算装置は、入力部３２と、外部知識２４と、入力特徴抽出部３４と、類似度計算部６０と、出力部３８とを含む。図８と図３とにおいて、同一のブロックには同一の参照符号および名称が与えられている。それらの機能も同一である。したがってここではそれらについての詳しい説明は繰返さない。

【００５２】類似度計算部６０は、ＣＰＵ６２と、メモリ６４と、内容でアドレス可能な連想メモリ６６とを含む。

【００５３】メモリ６４には、類似度計算のプログラムが格納されている。またメモリ６４には、各比較情報に対応する類似度を格納するエリアが設けられている。

【００５４】連想メモリ６６の各ワードには、１つの比較単語に対応した類語コード（比較コード）が予め格納されている。そして合計１０００個のワードに１０００個の比較コードが格納されている。

【００５５】連想メモリ６６は、マスクによる一致検索機能と部分並列書込機能とを有するものとする。マスクによる一致検索機能とは、ビット列の特定部分に対して一致検索を行なう機能をいう。部分並列書込機能とは、一致が検出された複数のデータの、特定のビットに並列にデータを書込む機能をいう。

【００５６】連想メモリ６６のデータの格納形式が図９に示される。図９を参照して、連想メモリ６６には比較コード１～１０００の１０００個のエリアが設けられる。各比較コードのためのエリア、たとえば比較コード１のエリアは全部で１５ビットからなる。このうち先頭から３ビットは類似度を求めるために使用されるエリアである。後半の１２ビットは比較コードを格納するためのエリアである。

【００５７】類似度を求めるためのエリアの３ビットの第１のビットは、入力コードと比較コードとの１００の位の一致／不一致を示すビットであり、２ビット目は入力コードと比較コードとの１０の位の一致／不一致を示すビットであり、３ビット目は入力コードと比較コードとの１の位の一致／不一致を示すビットである。

【００５８】比較コードを示す１２ビットはそれぞれ４ビットずつの３つの領域に分けられる。これら３つの領域は、図９に示されるように順に比較コードの１００の位を表わす４ビットと、比較コードの１０の位を表わす４ビットと、比較コードの１の位を表わす４ビットとである。

【００５９】なお、各比較コードエリアのうち最初の３ビットは、類似度計算に先立って初期値たとえばすべて０に設定されているものとする。

【００６０】以下、この第３の実施例の類似度計算装置の動作を説明する。キーボード、文字認識装置、音声認識装置などからなる入力部３２により入力単語が入力されると、その入力単語が入力特徴抽出部３４に与えられる。入力特徴抽出部３４は、与えられる入力単語に対応した類語コード（入力コード）を外部知識２４から抽出する。外部知識２４に格納されているデータは図４に示されたものと同様である。

【００６１】抽出された入力コードは類似度計算部６０に与えられる。類似度計算部６０は、次のようにして入力コードと１０００個の比較コードとの間の類似度計算を行ない、その結果を出力部３８に与える。

【００６２】類似度計算部６０では、連想メモリ６６を用いた次のような類似度計算が行なわれる。類似度の定義は、第１の実施例において図６を参照して説明したものと同じである。

【００６３】以下、入力単語として「取材」が入力された場合の、類似度計算部６０の動作を説明する。入力単語「取材」に対応する類語コード（入力コード）は図４に示されるように「４５７」である。この入力コード「４５７」が類似度計算部６０に与えられると、類似度計算部６０は次のように動作する。

【００６４】まず、図８のメモリ６４内に、図１１に示すように、図１０に示される連想メモリ内の比較コードの格納エリアと同様のデータ構造を有するエリアを設ける。このエリアは８行のエリア７０、７２、７４、７６、７８、８０、８２、８４からなる。第１行目のエリア７０の後半の１２ビットには、入力コードの「４５７」が格納される。このエリア７０の先頭の３ビットはすべて０である。この１行目のエリア７０を入力の検索コードと呼ぶ。

【００６５】第２行目のエリア７２には、入力の検索コードのうち、その１００の位を表わす４ビット以外をマスクしたデータが格納される。エリア７４には、入力の検索コードのうち１０の位を表わす４ビット以外をマスクしたデータが格納される。エリア７６には、入力の検索コードのうち１の位を表わす４ビット以外をマスクしたデータが格納される。エリア７８には、先頭の３ビットに“１１１”が格納され、それ以外のビットがマスクされたデータが格納される。エリア８０には、先頭の３ビットに“１１０”が格納され、それ以外のビットがマスクされたデータが格納される。エリア８２には、先頭の２ビットに“１０”が格納され、それ以外のビットがマスクされたデータが格納される。エリア８４には、先頭のビットに“０”が格納され、それ以外のビットがすべてマスクされたデータが格納される。以下、次に示すような手順に従って類似度計算が行なわれる。

【００６６】①　エリア７２に格納された、検索コード

の１００の位を表わす４ビット以外をマスクしたデータによる一致検索の命令と、一致した比較コードの１００の位の一致／不一致を示すビットに１を書込む命令とを順に連想メモリ６６に与える。

【００６７】②　エリア７４に格納された、検索コードの１０の位を表わす４ビット以外をマスクしたデータによる一致検索の命令と、一致した比較コードの１０の位の一致／不一致を示すビットに１を書込む命令とを順に連想メモリ６６に与える。

【００６８】③　エリア７６に格納された、検索コードの１の位を表わす４ビット以外をマスクしたデータによる一致検索の命令と、一致した比較コードの１の位の一致／不一致を示すビットに１を書込む命令とを順に連想メモリ６６に与える。

【００６９】上述の①～③の処理を行なった結果、連想メモリ６６上のデータは図１０に示されるようになる。すなわち比較コード１（４２６）と入力コード「４５７」とは１００の位のみが一致するために、図１０の比較コード１のエリアに示されるように１００の位の一致／不一致を示すビットのみが“１”となり、他の２ビットは“０”となる。比較コード２（１４９）と入力コードとは一致する桁がないために比較コード２の先頭の３ビットは“０００”となる。同じように比較コード３の先頭の３ビットは“１１０”となる。

【００７０】④　図１１のエリア７８に格納されたデータによる一致検索命令を連想メモリ６６に与え、一致した比較コードを検出する。そして、図８のメモリ６４上に予め準備されていた、各比較情報に対応する類似度を格納するエリアのうち、一致が検出された比較コードに対応するエリアに類似度として「３」を格納する（図６の条件（イ））。

【００７１】⑤　図１１のエリア８０に格納されたデータによる一致検索命令を連想メモリ６６に与え、一致した比較コードを検出する。そして、一致が検出された比較コードに対応するメモリ６４上のエリアに類似度として「２」を格納する（図６の条件（ロ））。

【００７２】⑥　図１１のエリア８２に格納されたデータによる一致検索命令を連想メモリ６６に与え、一致した比較コードを検出する。そして、１が検出された比較コードに対応するメモリ６４上のエリアに類似度として「１」を格納する（図６の条件（ハ））。

【００７３】⑦　図１１のエリア８４に格納されたデータによる一致検索命令を連想メモリ６６に与え、一致した比較コードを検出する。１が検出された比較コードに対応するメモリ６４上のエリアに類似度として「０」を与える。（図６の条件（ニ））。

【００７４】⑧　入力コードとすべての比較コードとの間の類似度を出力部３８に与える。上記した①～⑦で行なわれる一致検索命令および書込命令は連想メモリ上のすべての比較コードに対して並列に行なわれる。したがって第１の実施例と同様に、従来の逐次処理による類似度計算に比べて、単純計算で約１０００倍高速となり、類似度計算を高速に実現することができる。

【００７５】このように本発明に係る類似度計算装置を用いれば、入力情報と比較情報のそれぞれとの間の類似度を、外部知識から獲得した、入力情報に対応した入力特徴および比較情報に対応した比較特徴を用いて並列に行なわれる類似度計算によって求めることができる。特に、入力文と大量の用例（原文と訳文との対）とを比較して最も類似した用例を求めてそれを基に訳文を生成するタイプの機械翻訳システムや、大量の情報と入力情報との間の類似度を求める必要がある画像検索や文章検索などの情報検索において、入力情報に対応した特徴を用いて類似度を求める必要がある場合に、高速にそうした処理を行なうことが可能となる。

【００７６】

【発明の効果】以上のように請求項１ないし９に記載の類似度計算装置によれば、複数個の比較情報に対応して外部知識を用いて導出される比較情報の特徴のそれぞれと、入力情報に対応して導出される特徴との間で所定の類似度計算が並列に行なわれて類似度が導出される。したがって、このようにして比較特徴を用いて求める類似度計算において、従来の逐次処理と異なって計算速度の大幅な向上を実現できる類似度計算装置を提供できる。

【図面の簡単な説明】

【図１】本発明に係る類似度計算装置の概略の構成を示す図である。

【図２】類語辞書の分類体系の一部を示す図である。

【図３】本発明の第１の実施例の類似度計算装置の概略構成を示すブロック図である。

【図４】外部知識の内容を示す模式図である。

【図５】この発明の第１の実施例の類似度計算装置の類似度計算部内の計算機メモリに格納されている比較コードを示す図である。

【図６】この発明における入力コードと比較コードとの間の類似度の求め方を示す図である。

【図７】本発明の第２の実施例の類似度計算装置の概略構成を示すブロック図である。

【図８】本発明の第３の実施例の類似度計算装置の概略構成を示すブロック図である。

【図９】この発明の第３の実施例の類似度計算装置における連想メモリ上の比較コードの格納形式を示す図である。

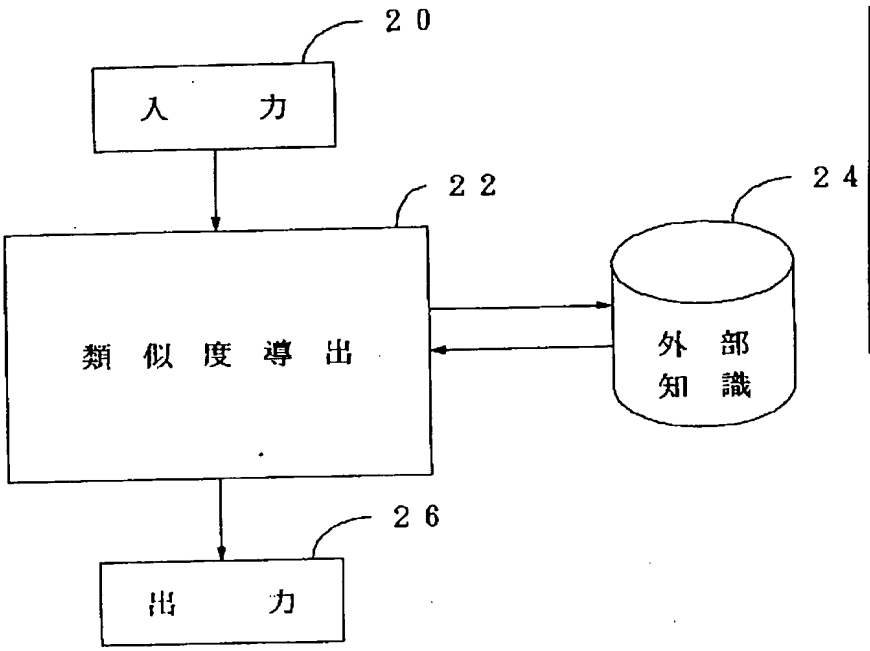
【図１０】この発明の第３の実施例における連想メモリ上の比較コードのデータを示す図である。

【図１１】この発明の第３の実施例の類似度計算装置の行なう類似度計算において使用する、メモリ上の検索のためのデータを示す図である。

【図１２】従来の類似度計算装置における類似度計算の手順を示す図である。

【符号の説明】
２４　外部知識
３２　入力部
３４　入力特徴抽出部
３６　類似度計算部
３８　出力部
５０　入出力管理部
５２　ネットワーク
６０　類似度計算部
６６　連想メモリ

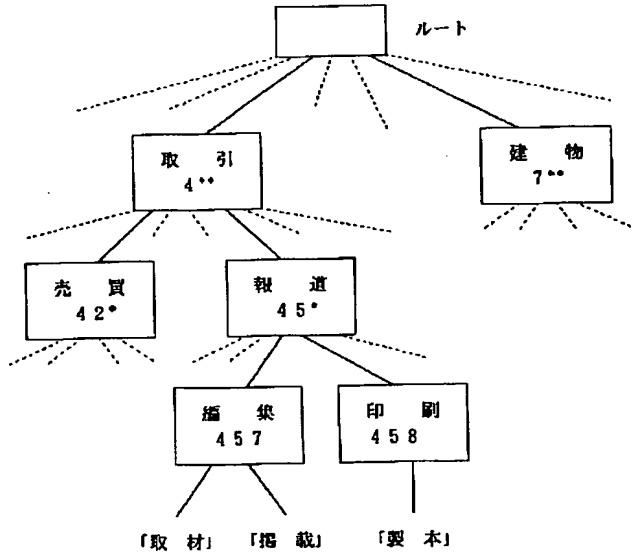
【図１】



【図４】

単語の見出し	類語コード
製　本	４５８
建　物	７３２
取　材	４５７
白　鳥	１４９
販　売	４２６
掲　載	４５７
：	：

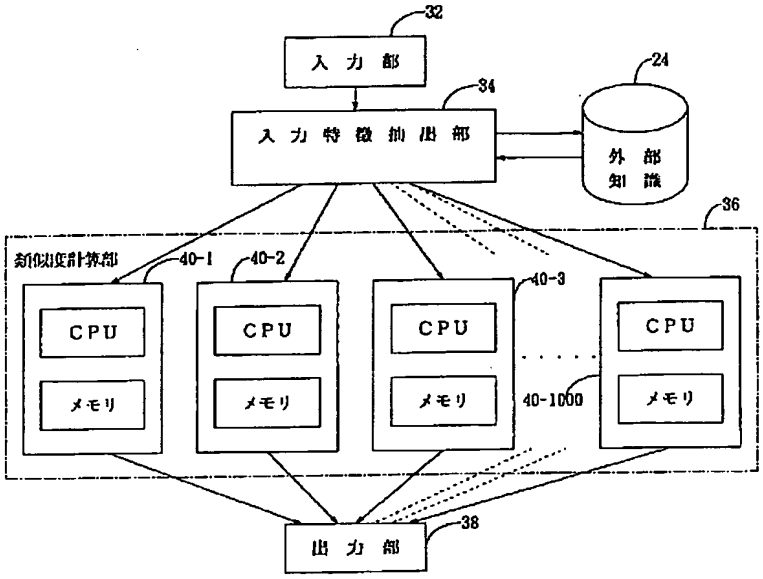
【図２】



【図６】

	入力コードと比較コードの条件	類似度	例（入力コード：比較コード）
(イ)	I_1 I_2 I_3 $=C_1$ C_2 C_3	3	４５７：４５７
(ロ)	I_1 I_2 $=C_1$ C_2　$I_3 \neq C_3$	2	４５７：４５８
(ハ)	$I_1 = C_1$　$I_2 \neq C_2$	1	４５７：４２６
(ニ)	$I_1 \neq C_1$	0	４５７：１４９

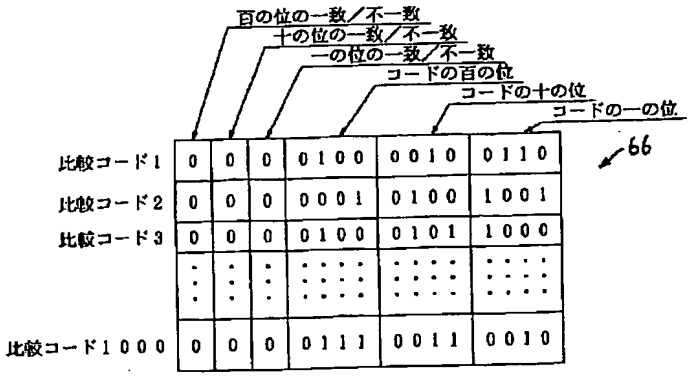
【図３】



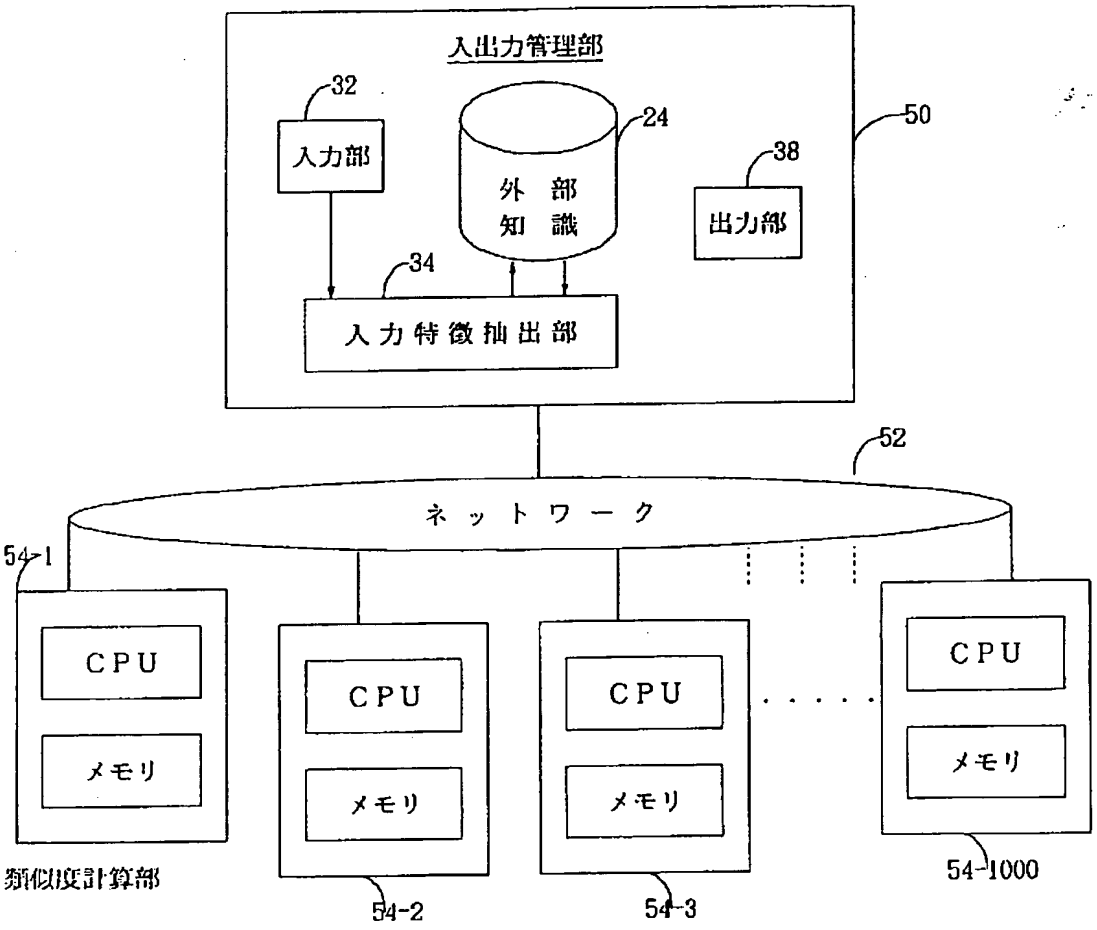
【図５】

類似度計算部コンピュータ	比較コード	
40－1	426	← 比較単語「販売」に対応する類語コード
40－2	149	← 比較単語「白鳥」に対応する類語コード
40－3	458	← 比較単語「製本」に対応する類語コード
・・・	・・・	
40－1000	732	← 比較単語「建物」に対応する類語コード

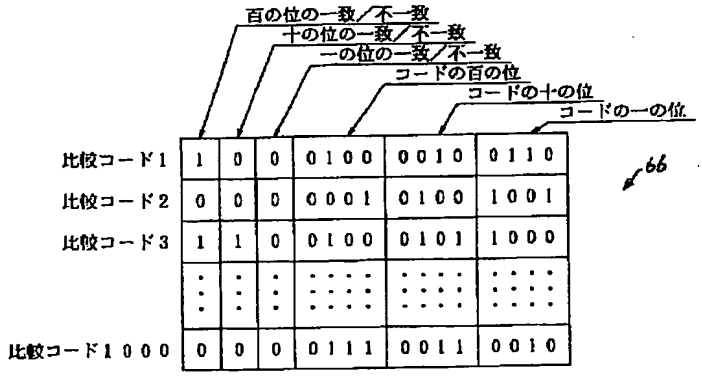
【図９】



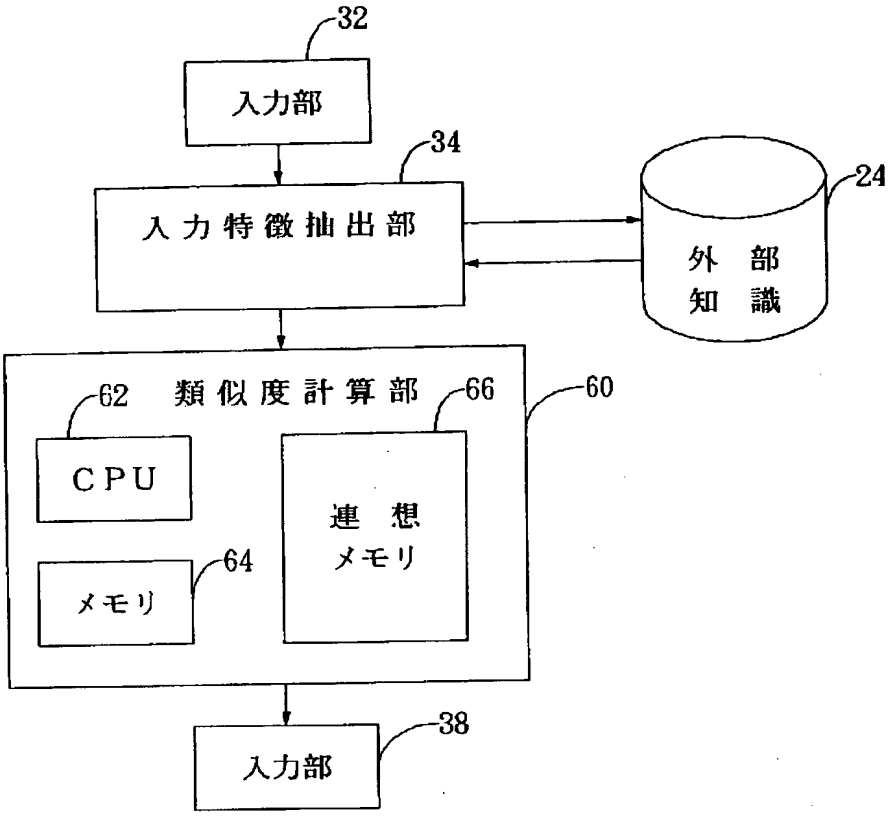
【図７】


入出力管理部
32
入力部
24
外部知識
38
出力部
50
34
入力特徴抽出部
52
ネットワーク
54-1
CPU
メモリ
類似度計算部
CPU
メモリ
54-2
CPU
メモリ
54-3
CPU
メモリ
54-1000

【図１０】


百の位の一致／不一致
十の位の一致／不一致
一の位の一致／不一致
コードの百の位
コードの十の位
コードの一の位
66
比較コード1 1 0 0 0100 0010 0110
比較コード2 0 0 0 0001 0100 1001
比較コード3 1 1 0 0100 0101 1000
比較コード1000 0 0 0 0111 0011 0010

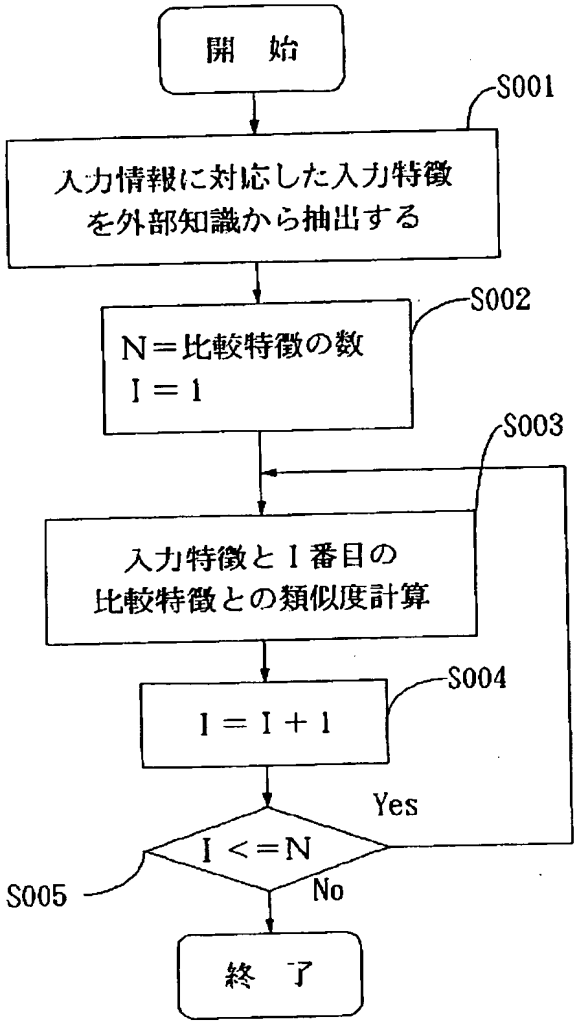
【図８】



【図１１】

	百の位の一致／不一致	十の位の一致／不一致	一の位の一致／不一致	コードの百の位	コードの十の位	コードの一の位
70: 入力の検索コード	0	0	0	0100	0101	0111
72～	X	X	X	0100	XXXX	XXXX
74～	X	X	X	XXXX	0101	XXXX
76～	X	X	X	XXXX	XXXX	0111
78～	1	1	1	XXXX	XXXX	XXXX
80～	1	1	0	XXXX	XXXX	XXXX
82～	1	0	X	XXXX	XXXX	XXXX
84～	0	X	X	XXXX	XXXX	XXXX

【図１２】



フロントページの続き

(72)発明者　隅田　英一郎
京都府相楽郡精華町大字乾谷小字三平谷５番地　株式会社エイ・ティ・アール自動翻訳電話研究所内

(72)発明者　飯田　仁
京都府相楽郡精華町大字乾谷小字三平谷５番地　株式会社エイ・ティ・アール自動翻訳電話研究所内